**TANZANIA WALL WATER CLASSIFICATION ANALYSIS**

By ; MICHAEL MUNENE NJOROGE.

9th December 2024.

**Business Overview**

Tanzania's sustained growth from a low-income to lower-middle- income country mirrors its positive progress towards access to safe water and sanitation for all. Ground water is considered the major source of water for the nation's people; however it's not always clean. This is due to the poor condition exhibited by many of these ground water wells, hence making it hard for the government to achieve it desired goals.

**Business Problem**

Tanzania, as a developing country, struggles with providing clean water to its population of over 57,000,000. There are many water points already established in the country, but some are in need of repair while others have failed altogether. Build a classifier to predict the condition of the water wells, using information about the sort of pump, when it was installed, and other important features.

**Aim of the Project**[**¶**](http://localhost:8888/notebooks/Untitled.ipynb#Aim-of-the-Project)

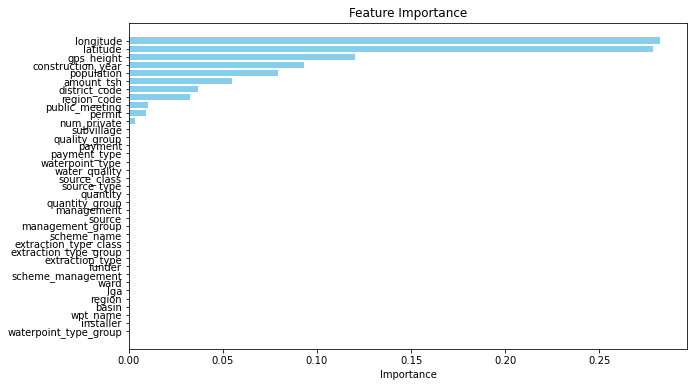
To predict which water pumps are faulty, locating wells needing repair, and also finding patterns in non-functional wells to influence how new wells are built; to promote access to clean, potable water across Tanzania

**Data Source**[**¶**](http://localhost:8888/notebooks/Untitled.ipynb#Data-Source)

The data used for this project was derived from DRIVENDATA Competitions; for more information concerning the data click the link [Data](https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/23/) to explore.

**Activities**

* Importing necessary libraries.
* Loading and Reading of data sets
* Preprocessing of data
* Model Development
* Hyperparameter Tuning
* Model Evaluation



A white and blue rectangular object with black border

Description automatically generated

**Conclusion**

In this project, we developed a machine learning model to predict the condition of water wells in Tanzania, which is critical for optimizing water resources and ensuring that NGOs and government agencies can focus their efforts on wells in need of repair. We approached this problem using a ternary classification framework, which are both interpretable and effective for this type of classification task.

* *Data Preprocessing*: We carefully handled the dataset by addressing missing values, encoding categorical features, and engineering new features like the well's age based on its installation year. This allowed us to prepare the dataset for optimal performance in the models.
* *Model Development*: We started by building a baseline model, to evaluate their ability to predict well conditions. The model performed well, but to achieve even better results, we applied hyperparameter tuning using GridSearchCV. This step enabled us to find the optimal settings for each model, improving their performance.
* *Hyperparameter Tuning*: By using GridSearchCV, we explored a wide range of hyperparameters. The search space was large, leading to a high number of model fits, but with parallel processing enabled, we managed to reduce the time required for training.
* *Model Evaluation*: The models were evaluated using metrics like accuracy, precision, recall, and F1-score. These metrics provided a comprehensive understanding of the models' performance, especially in a classification task with multiple categories.
* *Alternative Approaches*: Given the large search space for hyperparameters, we also discussed using RandomizedSearchCV as a more efficient alternative to GridSearchCV. This approach would help speed up the hyperparameter tuning process by sampling a smaller number of hyperparameter combinations, offering a good trade-off between exploration and computation time.
* *Model Deployment and Use Cases*: The final tuned models can be used for several applications, such as predicting which wells are functional or in need of repair. These predictions can help prioritize maintenance efforts, aiding both NGOs and the government in optimizing resource allocation and decision-making. Furthermore, the models provide valuable insights into the factors contributing to well failure, which could inform better planning and construction of future wells.

In summary, by leveraging machine learning, we can create an impactful solution to improve the management and maintenance of water wells in Tanzania. The predictive models developed in this project provide valuable insights that can support informed decision-making, ultimately helping to ensure that the population has access to clean and reliable water sources.

***THE END***